Milestone 3: Complete Search System

Search Engine Implementation Report

Executive Summary

This report documents the complete search engine implementation for Milestone 3, including optimizations for disk-based access, enhanced ranking algorithms, and comprehensive testing with 30 diverse queries. The search engine meets all M3 requirements including response time (≤300ms), memory efficiency, and improved retrieval effectiveness.

Key Improvements

- Disk-based index access without loading entire index into memory
- Enhanced TF-IDF with sublinear scaling and smoothed IDF
- Important word boosting (2x for bold, headings, titles)
- Query length normalization to prevent bias
- Complete match bonus for documents containing all query terms
- Efficient posting list intersection algorithm
- LRU caching for frequently accessed postings

Test Queries

We tested the search engine with 30 queries divided into three categories: 10 good-performing queries (specific, clear intent), 10 poor-performing queries (too general or common words), and 10 challenging queries (edge cases).

Good Performing Queries

These queries are specific and should perform well. They test the search engine's ability to find relevant content for clear, well-formed queries.

Query	Results	Time (ms)	Top Result
cristina lopes	10	33.03	https://www.informatics.uci.edu/explore/faculty-pr...
machine learning	10	34.57	https://www.cs.uci.edu/singh-to-deliver-machine-le...
ACM	10	14.39	https://www.informatics.uci.edu/two-ics-professors...
master of software engineering	10	59.60	https://www.informatics.uci.edu/grad/mswe/...
informatics department	10	25.94	https://www.informatics.uci.edu/explore/department...

Poor Performing Queries (Before Improvements)

These queries initially performed poorly due to being too general or containing very common words. We implemented general heuristics to improve their performance.

Query	Results	Time (ms)	Issue
the	10	18.54	too common word
a	10	14.66	single letter
and	10	20.34	common word
computer	10	1.07	too general
student	10	18.59	too general

Improvements Implemented

1. Disk-Based Index Access

Implemented term-specific JSON extraction to read only needed postings from disk. Uses regex and brace matching to extract term data without loading the entire 14MB index. Memory footprint reduced from 14MB to < 1MB.

2. Enhanced TF-IDF Calculation

Improved ranking with sublinear TF scaling (log(1+tf)) and smoothed IDF. Increased important word boosting from 1.5x to 2.0x. Added query length normalization to prevent bias toward longer queries.

3. Complete Match Bonus

Added 15% boost for documents containing all query terms, improving results for multi-term queries.

4. Efficient Algorithms

Optimized posting list intersection by starting with smallest lists. Implemented LRU caching for frequently accessed postings. Removed duplicate terms in query processing.

Performance Metrics

Metric	Value
Total Queries	30
Average Response Time	22.58 ms
Min Response Time	0.69 ms
Max Response Time	59.60 ms
Queries < 300ms	30/30 (100.0%)
Queries < 100ms	30/30 (100.0%)
Good Queries Avg	35.63 ms
Poor Queries Avg	9.02 ms

General Heuristics

All improvements use general heuristics that apply to all queries, not query-specific fixes:

- Sublinear TF scaling applies to all terms
- Smoothed IDF applies to all terms
- Important word boosting applies to all important terms
- Query length normalization applies to all queries
- Complete match bonus applies to all multi-term queries
- Efficient intersection applies to all boolean AND queries

Conclusion

The M3 search engine successfully meets all requirements: response time ≤ 300ms, memory-efficient disk-based access, and improved retrieval effectiveness. All improvements are general heuristics that enhance performance across diverse query types without query-specific optimizations.